Unimore @ ImageCLEF 2013: Scalable Concept Image Annotation

Costantino Grana, Giuseppe Serra, Marco Manfredi, Rita Cucchiara, Federica Mandreoli and Riccardo Martoglia

University of Modena and Reggio Emilia

Outline

- ImageCLEF2013 complexity and possible approaches
- Our solution
  - Image Description
  - Text analysis
  - Enhancing the Training Set
- Experimental Results
- Conclusions

ImageCLEF2013

- **Annotation Task:**
  - 250000 Training Images
  - 95 (develop), 116 (test) concepts to be identified
  - A lot of label **Noise** inside the training set, due to the automatic label extraction from websites

ImageCLEF2013

- **Possible Approaches:**
  1. Given a query image, find visually similar images in the training set, and from them extract the query concepts
     - The baseline proposed by the organizers belong to this group of strategies

  2. Use the training set text annotations to build a classifier for each concept. Use these classifiers to annotate the query.
     This strategy **outperformed** the first baseline in a preliminary experiment (using Bag Of Words on CSIFT features), so **we further expanded this approach.**

Training Images Annotation

- The annotation of the training images is done exploiting the *scofeat* file given by the organizers.
  In the scofeat file, **each image** is associated with a **list of words**, automatically extracted from the web page in which the image was found. Each word has a **score** related to its relevance.

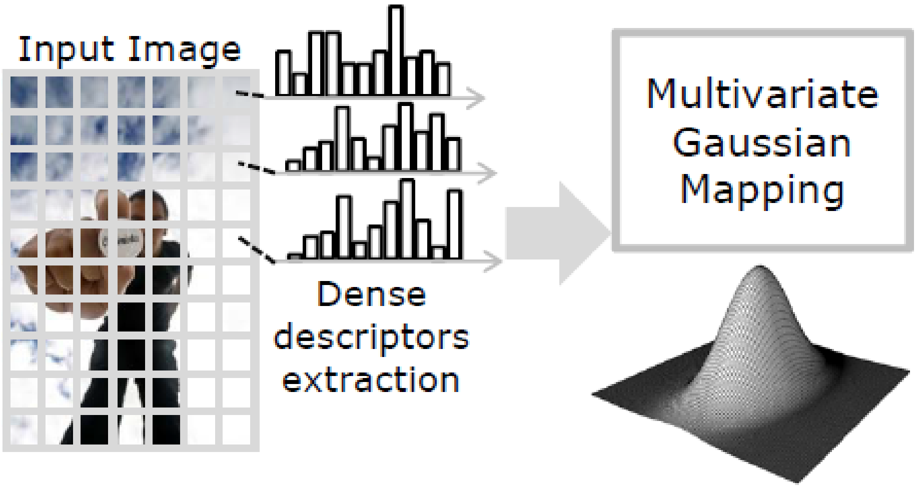
1080p	261k	3d	aircraft	all	and
animal	animals	apple	are	as	
background	backgrounds	big	bikes	brands	
butterfly	by	cars	cartoons	cat	
celebrities	choose	click	commercial	computers	
copyrighted	custom	definition	desktop	deutsch	dogs
downloads	email	english	español	europe	
facebook	fantom	female	**flowers**	foods	
forbidden	français	free	from	full	funny
games	graphics	hd	high	holiday	home
humor	image	**insects**	ipad's	iphone	is
italiano	keywords	kitten	languages	life	links
location	mac	men	military	miscellaneous	
most	motors	movies	music	papilio	pc
polytes	popular	random	resolution	resolutions	right
search	set	share	site	smartphone	
standart	story	the	to	wallpaper	
wallpapers	widescreen	windows	xp		

Our solution

1. Improve the **Visual Features** extracted from images, starting from the SIFT variants given by the organizers.
   Instead of relying on the BoW model we propose to describe the local features as a **Multivariate Gaussian Distribution**, with full rank covariance matrix
2. Improve **Textual Annotations,** relying on stopwords removal, stemming and on **WordNet** to build a context around each concept used for further **analysis**
3. Improve the training set, crawling from images using **Google Image Search**
4. **Late fusion** approach to fuse various sources of information
5. **Online Learning** using a SGD solver.

Visual Features

1. Extract local features (e.g. SIFT) from images on a regular grid
2. Describe the local features distribution with a **Multivariate Gaussian Distribution,** thus obtaining a fixed length descriptor, composed of the mean and the full rank covariance matrix.



Multivariate Gaussian Descriptor

- The set of local features of an image is modeled with their **mean** vector and **covariance** matrix:

$$\mathcal{N}(\mathbf{f};\mathbf{m},\mathbf{C}) = \frac{1}{|2\pi\mathbf{C}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{f}-\mathbf{m})^T\mathbf{C}^{-1}(\mathbf{f}-\mathbf{m})}$$

- The covariance matrix does not lie in a vector space (we can not compute dot product), in fact, it lies on a Riemannian manifold. To work with **linear classifiers**, we have to project it on a **Euclidean** space, as previously proposed by Tuzel *et al.*[1]

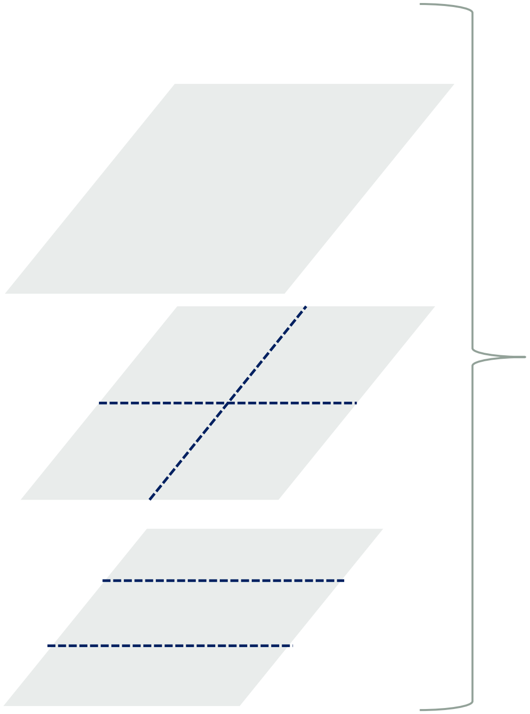
1. Tuzel, O., Porikli, F., Meer, P.: Pedestrian Detection via Classication on Riemannian Manifolds. IEEE T. Pattern Analysis and Machine Intelligence

Multivariate Gaussian Descriptor

- Each set of local features is thus described with the **concatenation** of the mean and the covariance matrix;
- when SIFT are used as local features, the mean is a **128** dimensional vector and the projected covariance matrix is (128*128+128)/2 = **8256** dimensional.
  Thus leading to a **8384**-dimensional feature vector.

Spatial Pyramid

- We partitioned the image into 1X1, 2X2, 1X3 regions, following the Spatial Pyramid approach of Lazebnik *et al.*[2]



We obtain 8 regions, each of them described with a multivariate Gaussian descriptor.
The image representation is the concatenation of the regions' description, obtaining a:
8384 X 8 = **67072** feature vector

2. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in CVPR 2006

Text Analysis

- Given the list of **concepts of interest** proposed by the organizers, the goal is to retrieve a **relevant** set of images in the ImageCLEF training set, exploiting only the textual content of the web pages that referenced the images
- The concepts are expressed as **WordNet** synsets, removing any label ambiguity
- The set of relevant images must be:
  - Sufficiently large to perform training
  - As relevant as possible

Text Analysis

Using the *scofeat* file and the relative webpages, the main steps are:

1. Stopword removal and Stemming – to clean the labels
2. Extraction and analysis of the titles of the webpages
3. Extraction of synonyms and hyponyms, enlarging the training set
   - Only synonyms and hyponyms with a single sense in WordNet are selected, to avoid noise in the results
4. Filtering and refinement:
   1. *scofeat* score threshold, to maintain only relevant words
   2. negative context generation, to exclude words related to other senses of the same word in WordNet

Enlarging the Training Set

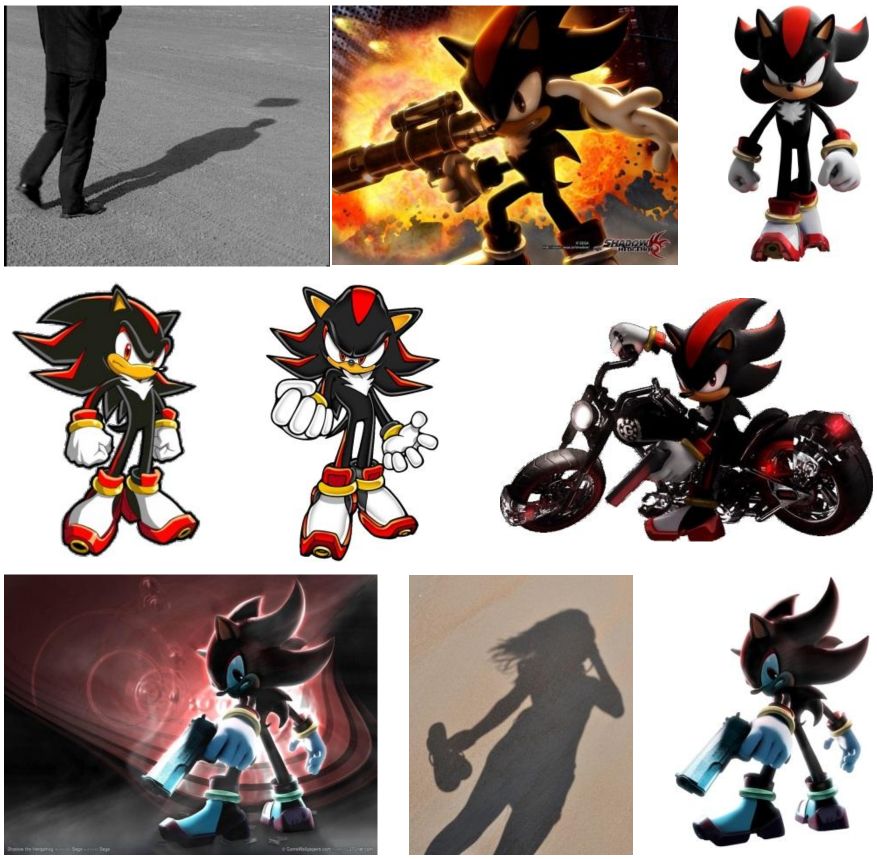
- The training set is **big**, but we want to make it even **larger**, adding useful information from the web;
- we choose **Google Images Search** to automatically download a large amount of images of the 116 concepts of the competition;
- **no manual filtering was applied to the images;**
- **103958** images were downloaded, by querying 1000 images per concept, and filtering out broken files
- We called this additional training set **Google100K**

Google100K

Querying: “**airplane**”

useful…

Querying: “**shadow**”



…harmful

Learning

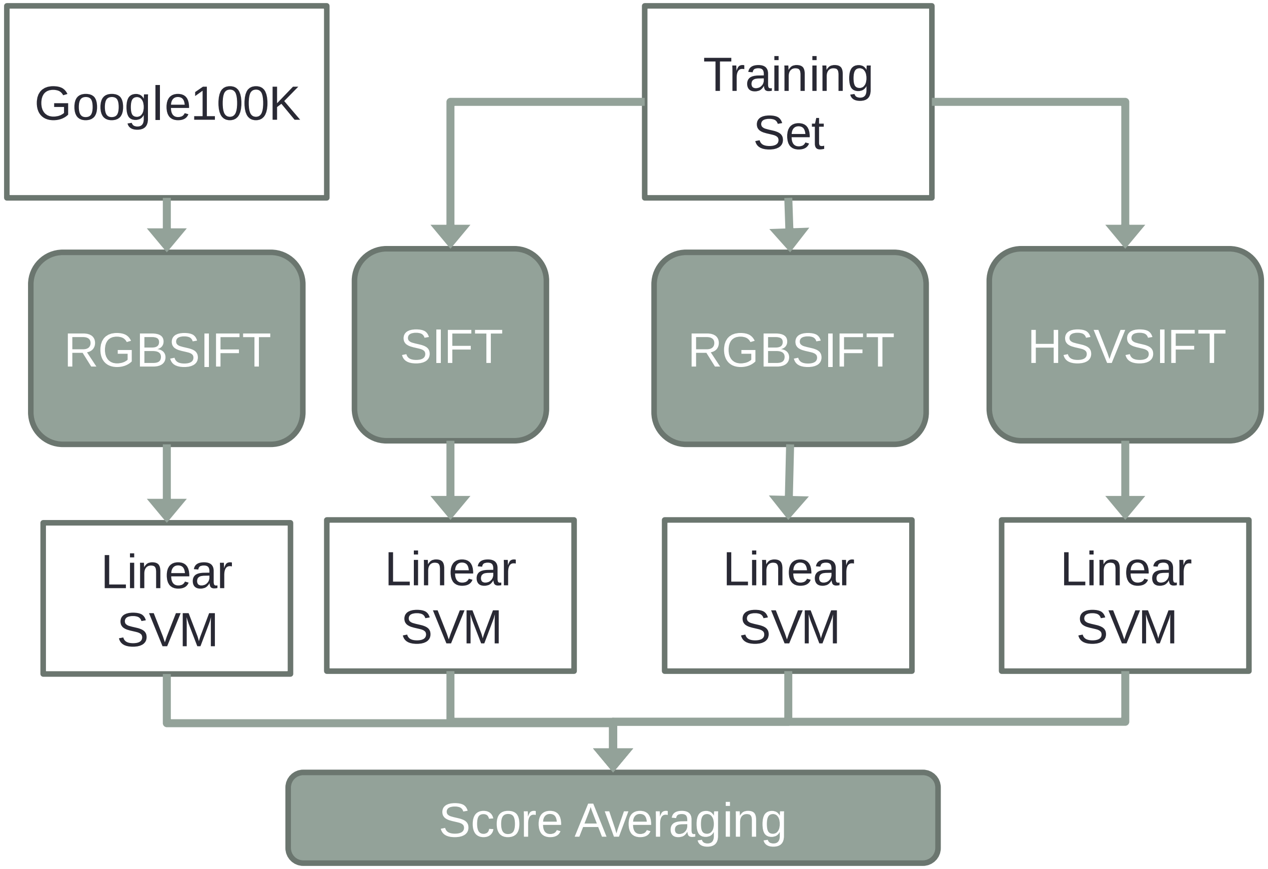
- Once we have a visual description of images, and a text annotation, we can learn a set of **1-vs-all** linear SVMs**;**
- For each testing image, a list of concept must be provided as output, sorted from the most relevant to the least relevant. We sorted the concepts using the scores of the SVMs;

Late Fusion

- We adopt a simple late fusion approach to:
  - Exploit different local descriptors, such as SIFT, HSVSIFT, RGBSIFT, OpponentSIFT;
  - Mixing the training set given by the organizers and the Google100K training set;
  - Mixing various text analysis approaches

- The late fusion approach consists in **averaging** the scores of the classifiers learned using different strategies listed above.

Late Fusion - Example



Training Set Complexity

- Each SVM is trained with approximately 250000 samples, with highly **unbalanced** data, having few positive samples and a large amount of negative ones;
- the training set is **noisy,** that means that a lot of incorrect images are associated to a concept;
- this complicates the training phase, leading to testing scores **biased towards the negative samples;**
- obtaining binary decisions from this scores is difficult, because they are all negative;
- for this reason we **optimize the SVM bias** to maximize the F-measure on the training set;
- thresholding the score to zero (as usual) gives us the decisions to compute F-measure

Online Learning

- Given the very high dimensional feature vector (**67072** using SIFT, **201216** using SIFT color variants), and the number of training images (250000+) an **online** solver is chosen, instead of a **batch** solver.
- The online solver takes one example at a time, and thus does not need to load the entire training set in memory;
- We chose the **Stochastic Gradient Descent** solver (SGD);

Experimental results

- 6 runs have been submitted to ImageCLEF2013:
  - The simplest run consists in using HSVSIFT and RGBSIFT as local descriptors, the plain scofeat file is used without any further text analysis; a late fusion approach is used;
  - The other runs add Google100K training set, text analysis and all the SIFT variations listed previously

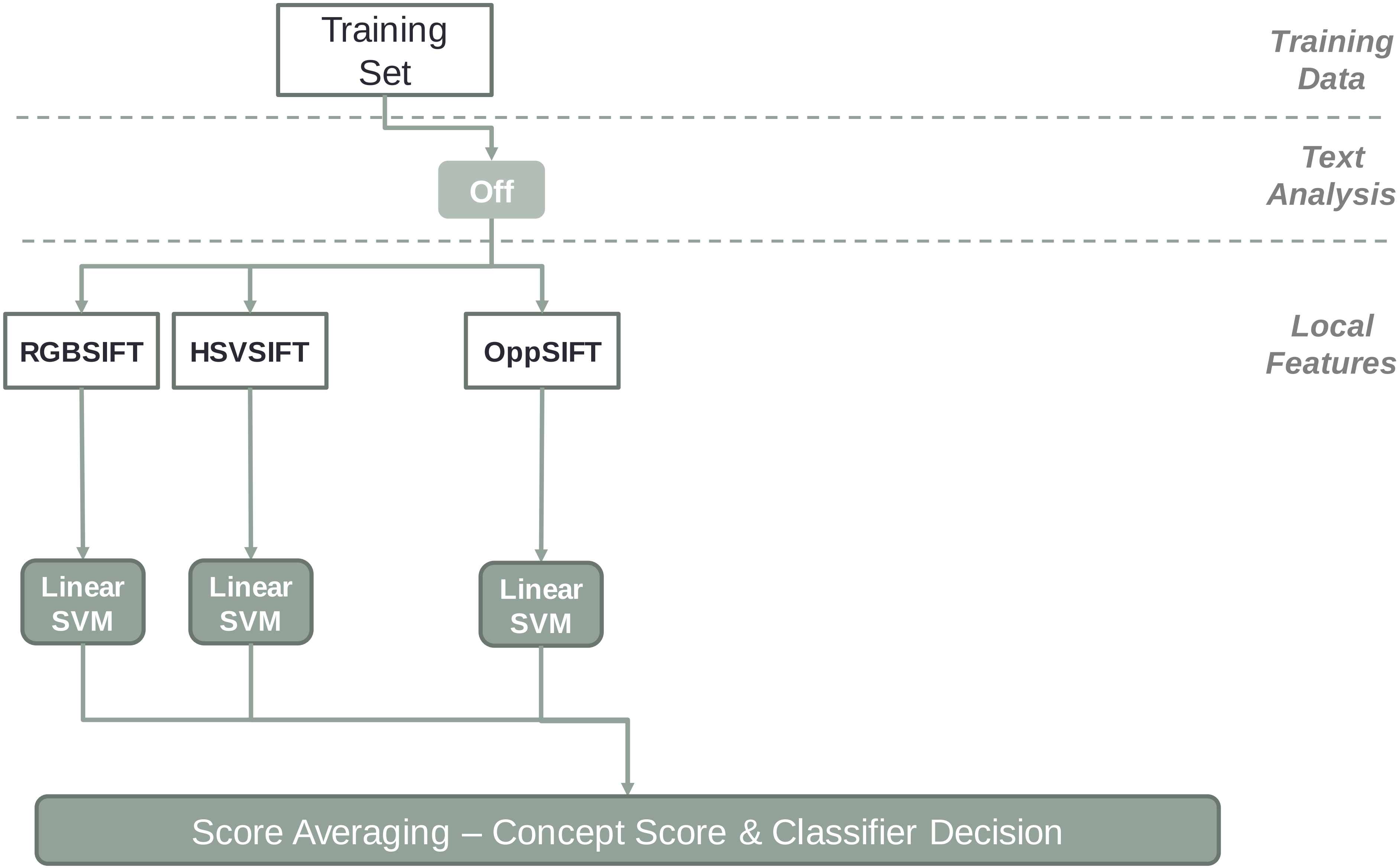
Experimental results

- Results on the **Test set:**

	MF-samples	MF-concepts	MAP-Samples
baseline_rand	4.6	3.6	8.7
baseline_sift	15.9	11.0	21.0
UNIMORE_1	31.1	32.0	36.7
UNIMORE_2	27.5	33.1	44.1
UNIMORE_3	23.1	31.5	41.9
UNIMORE_4	24.1	29.5	36.2
UNIMORE_5	31.5	31.9	45.6
UNIMORE_6	31.1	32.0	44.1

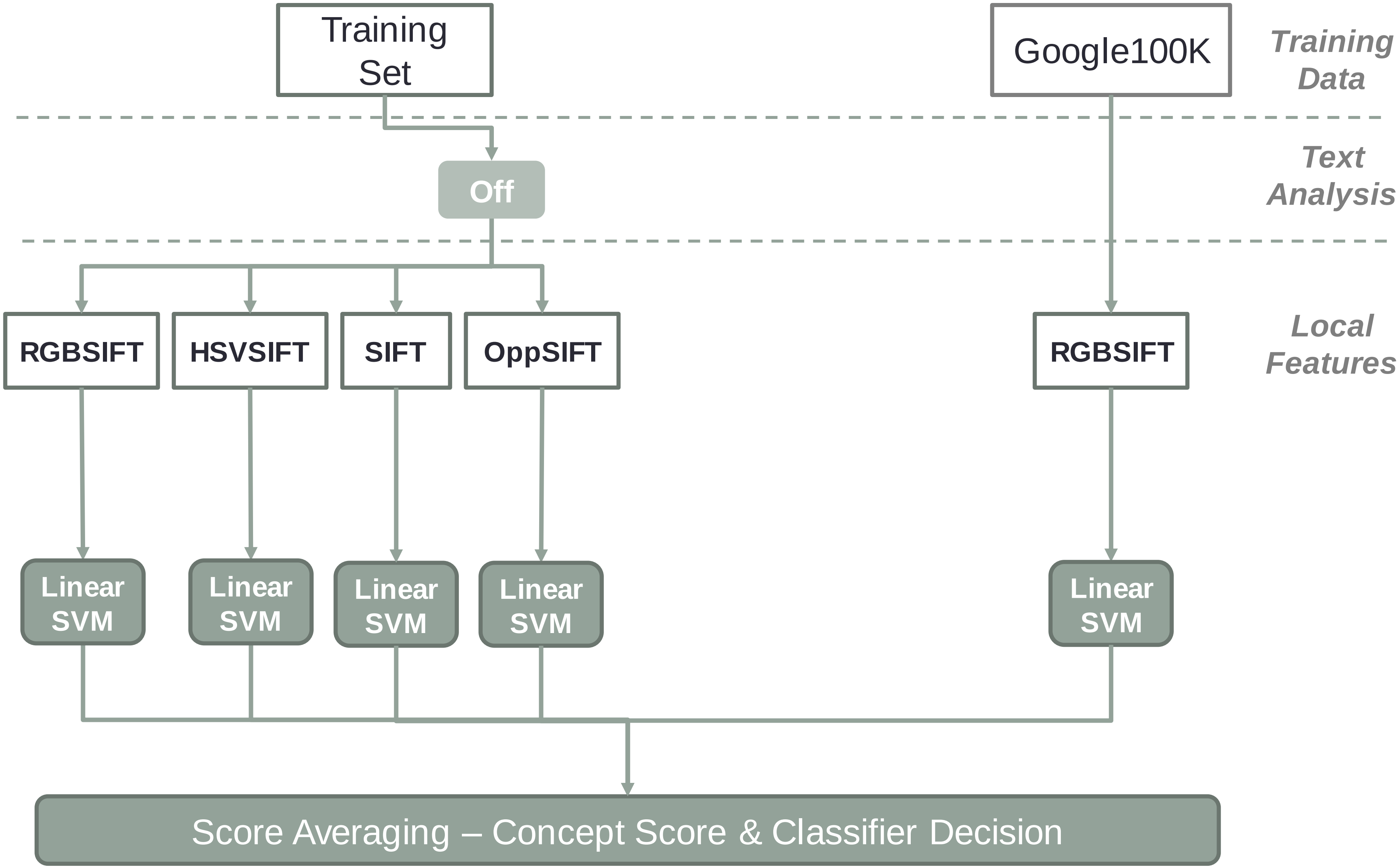
Run 4

F-sample	F-concept	MAP
24.1	29.5	36.2



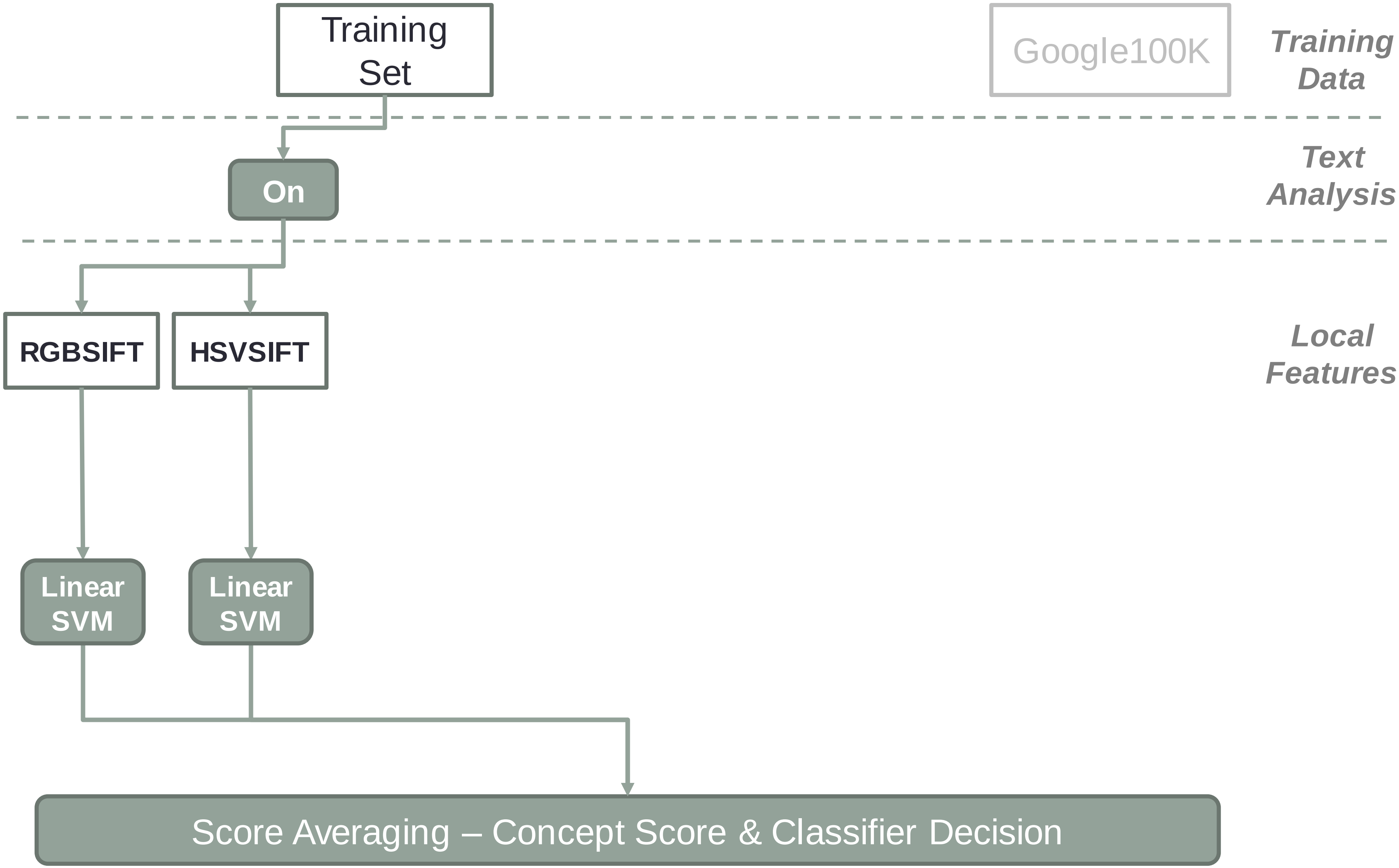
Run 3

F-sample	F-concept	MAP
23.1	31.5	41.9



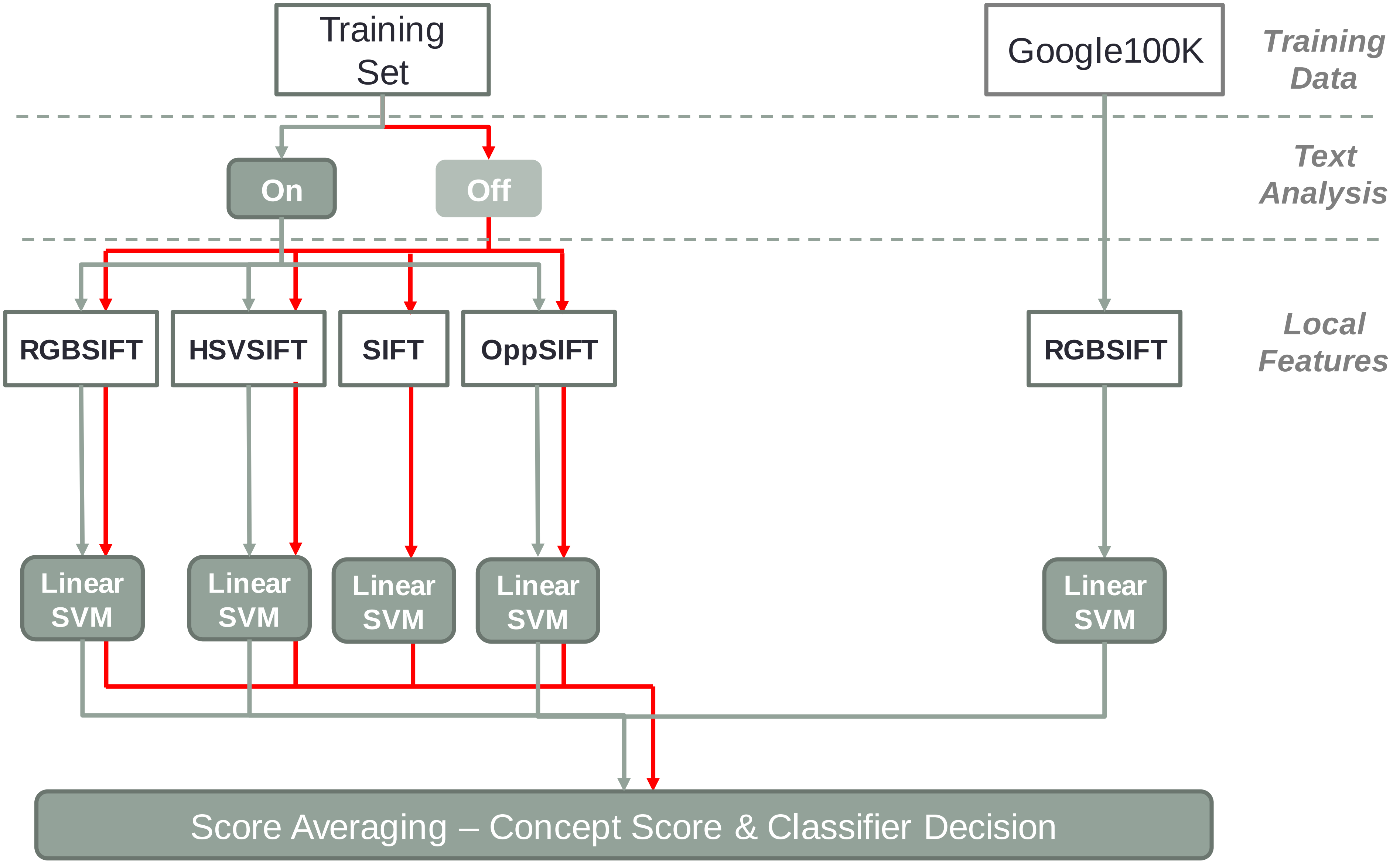
Run 1

F-sample	F-concept	MAP
31.1	32.0	36.7



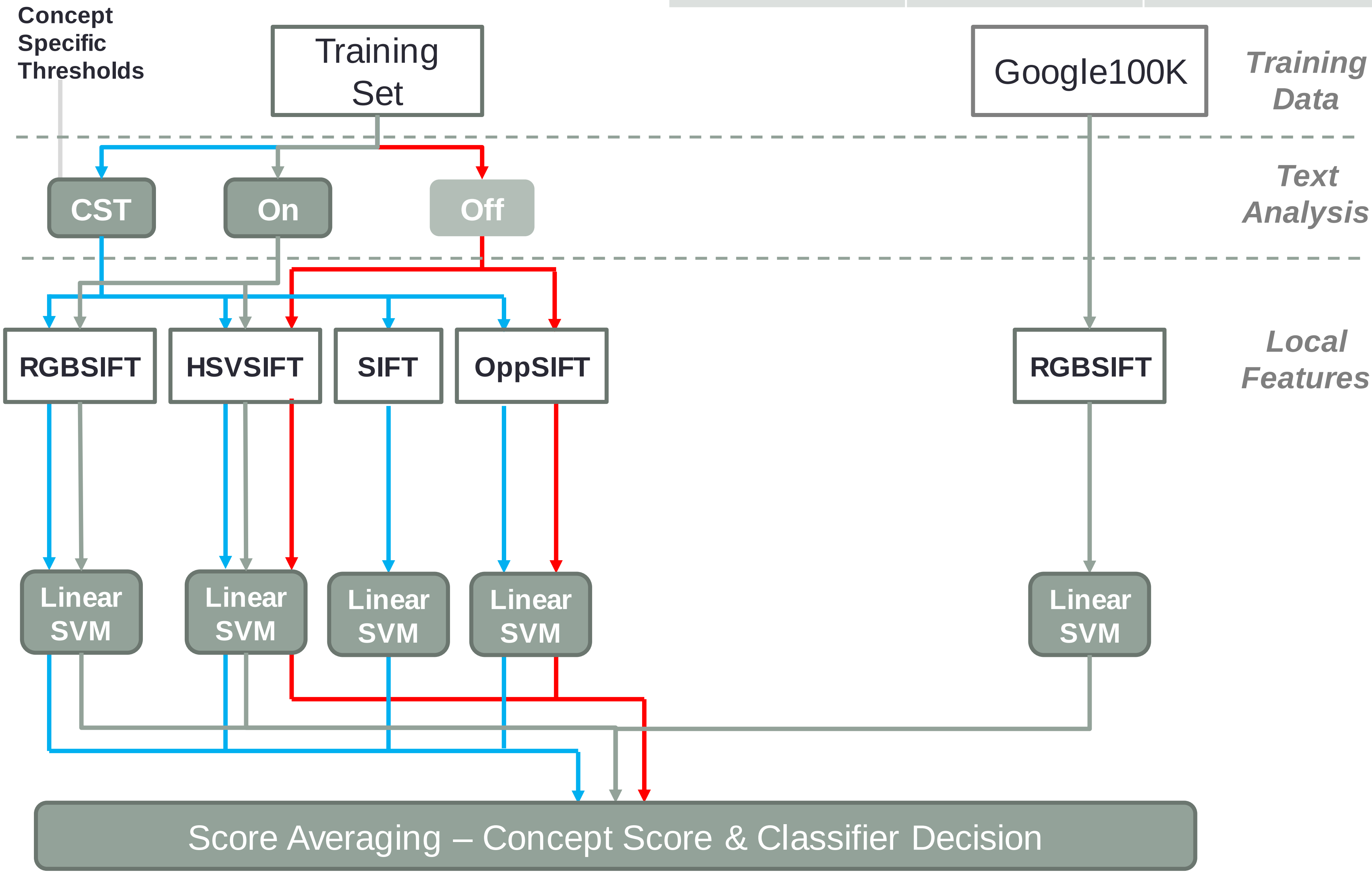
Run 2

F-sample	F-concept	MAP
27.5	33.1	44.1



Run 5

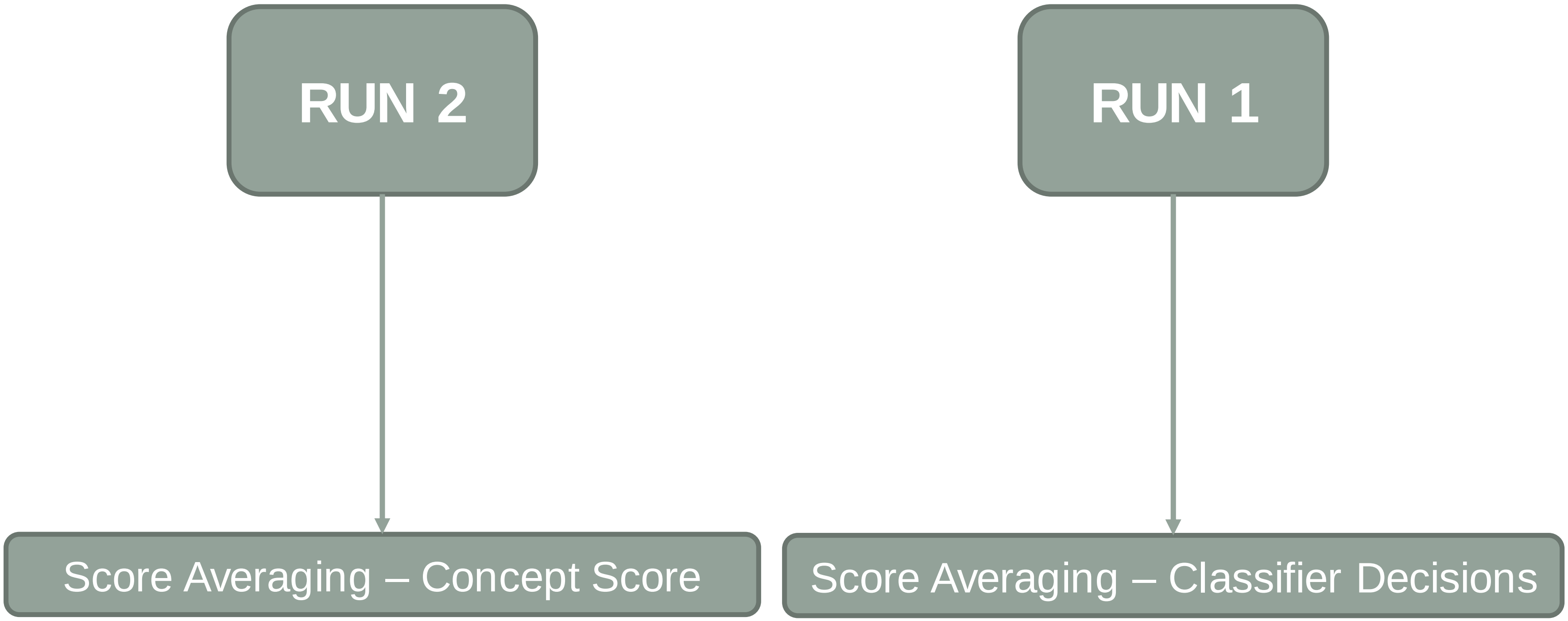
F-sample	F-concept	MAP
31.5	31.9	45.6



Run 6

F-sample	F-concept	MAP
31.1	32.0	44.1

- The Run 6 is a balanced run, in which we used the approach of Run 1 to compute the binary decisions, and the approach of Run 2 to compute the score for each concept.



Experimental results

- The Mean Average Precision, that measures the order of the concepts for each test sample, is greatly improved using late fusion on multiple approaches
- The F-measure, instead, is not substantially affected

MF-concepts	MAP-Samples
32.0	36.7
33.1	44.1
31.5	41.9
29.5	36.2
31.9	45.6
32.0	44.1

Comments

- Modeling local features with a Multivariate Gaussian is effective and leads to state-of-the-art results;
- using several SIFT variations in a late fusion approach is useful and enhance considerably the performance;
- text analysis helps to clear the training set;
- retrieving images from Google gives 4-5 percentage points of MAP

Conclusions

- We presented a new image descriptor that encodes local features, densely extracted from a region, as a Multivariate Gaussian Distribution;
- a new textual information processing strategy is also presented to cope with the high level of noise of the training data;
- to deal with the large-scale nature of this task, we use an online linear SVM classifier based on the Stochastic Gradient Descent algorithm;
- the proposed approach obtained the **best MAP** over the testing set.